

# Оценивание вероятности ошибочной классификации

Неделько В. М.

Институт математики СО РАН, г. Новосибирск  
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».  
Лекция 8.

## «Парадокс конвертов»

Игроку предлагается выбрать один из двух одинаковых на вид запечатанных конвертов с деньгами, причём известно, что сумма в одном из них в 10 раз больше, чем в другом. При этом игроку разрешается вскрыть один конверт, после чего решить, забрать его или оставшийся запечатанным.

Вероятностная модель: ведущий в конверт  $A$  кладёт сумму  $x$ , в конверт  $B$  помещает равновероятно  $10x$  или  $\frac{x}{10}$ .

## Решение «парадокса»

Математическое ожидание выигрыша  $f$  есть

$$E_A f = x, \quad E_B f = \frac{1}{2} \cdot 10x + \frac{1}{2} \cdot \frac{x}{10} = 5,05x.$$

Если не знаем, где какой конверт, и берём наугад

$$E f = \frac{E_A f + E_B f}{2} = 3,025x.$$

При условии, что в первом конверте обнаружили  $y$

$$E_x f = E_B = 5,05x, \quad E_{10x} f = E_{0,1x} f = x.$$

Величина  $E_y$  при других значениях  $y$  не определена.

В задаче мы не знаем  $E_y$ , поскольку не знаем, как  $y$  соотносится с  $x$ .

## Модельный пример

Известно, что в урне белые и чёрные шары. Извлекли 10 шаров, все оказались белыми. Какой прогноз о цвете следующего шара?

Пусть  $p$  – вероятность чёрного шара

$$P_p(M) = C_N^M p^M \cdot (1 - p)^{N-M}, \quad P_p(0) = (1 - p)^N.$$

Положив  $P_p(0) = \alpha$ , имеем  $p = 1 - \alpha^{\frac{1}{N}} = 1 - e^{\frac{\ln \alpha}{N}}$ .

При  $\alpha = 0,1$  и  $N = 10$  получим  $p \approx 0,2$ .

# Доверительный интервал

Односторонний интервал  $[0, \hat{p}]$

$$\sum_{i=0}^M C_N^i \hat{p}^i \cdot (1 - \hat{p})^{N-i} = \alpha.$$

Двусторонний интервал  $[p_1, p_2]$

$$\sum_{i=0}^M C_N^i p_2^i \cdot (1 - p_2)^{N-i} = \sum_{i=M}^N C_N^i p_1^i \cdot (1 - p_1)^{N-i} = \frac{\alpha}{2}.$$

# Байесовский подход

Положим равномерное  $\varphi(p) \equiv 1$ .

Формула Байеса

$$\varphi(p | M) = P(M | p) \frac{\varphi(p)}{P(M)}.$$

Используя нормировку, получаем

$$\varphi(p | M) = (N + 1) C_N^M p^M \cdot (1 - p)^{N-M}.$$

Можем вычислить

$$E_M p = \int_0^1 p \varphi(p | M) dp = \frac{M + 1}{N + 2}.$$

## Усреднение по доверительной вероятности

Считаем  $\eta(\hat{p}) = 1 - \alpha(\hat{p})$  функцией распределения.  
Можно усреднить

$$\hat{E}_{Mp} = \int_0^1 \hat{p} d\eta(\hat{p}) = \frac{M+1}{N+1}.$$

Если нельзя, но очень хочется, то — можно.

# Нормальное приближение

Обозначим  $\nu = \frac{M}{N}$ .

Имеет место оценка

$$P(|\nu - p| > \varepsilon) \approx 1 - \Phi\left(\frac{\varepsilon\sqrt{N}}{\sqrt{p(1-p)}}\right) \leq 1 - \Phi(2\varepsilon\sqrt{N}) \leq e^{-2\varepsilon^2 N},$$

где  $\Phi(x) = 2 \int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$  – функция Лапласа.



# Случай конечного множества решающих правил

Пусть  $L$  – число решающих функций  $\lambda$ .

Имеем

$$\mathbf{P}(\forall \lambda, |\nu_\lambda - p_\lambda| > \varepsilon) \leq L e^{-2\varepsilon^2 N} = e^{\ln L - 2\varepsilon^2 N}.$$

Использовали  $\mathbf{P}(\sum A_i) \leq \sum \mathbf{P}(A_i)$ .

Получаем

$$\frac{\ln L}{2N} \approx \varepsilon^2.$$

# Случай бесконечного множества решающих правил

Можно считать различными только те правила, которые по-разному классифицируют выборку.

Если выборка классифицируется всеми  $2^N$  возможными способами, то класс правил имеет бесконечную ёмкость или сложность.

## Замечания

- Отклонение эмпирического риска от риска зависит от сложности метода.
- Оценки Вапника-Червоненкиса представляют только теоретический интерес, поскольку сильно завышены.
- Следует различать сложность метода и сложность класса правил.
- Методы «бесконечной» сложности также могут давать хорошие решения.